Appendix E. **Data Quality**

Two principal indicators of the quality of data collected in household surveys are the magnitude of imputed and modified responses, and the accuracy of the responses that are provided. Another source for data quality is through comparisons to administrative estimates. This appendix provides a review of the data quality of the Wave 5 School Enrollment and Financing topical module from the Survey of Income and Program Participation (SIPP). The data are discussed in the context of imputation rates, comparisons to other sources, and overall reasonableness of the data, as well as some of the problems encountered in collecting the data.

IMPUTATION RATES

Imputed responses refer either to missing responses for specific questions or "items" in the questionnaire, or to responses that were rejected in the editing procedure because they were improbable or inconsistent. Persons may not respond for a variety of reasons, and nonresponse may occur for the entire topical module or only for chosen items.

The estimates shown in this report are produced after all items have been edited and imputed wherever necessary. Missing or inconsistent responses to specific questions are assigned a value in the imputation phase of the data processing operation. The procedure used to assign or impute responses for missing or inconsistent data is referred to as the "hot deck" imputation method. This process assigns item values reported in the survey by respondents to those who do not respond. The respondent from whom the value is taken is called the "donor." Values from donors are assigned by controlling for demographic and labor force data available for both donors and nonrespondents.

Imputation rates for some of the major items in this report are shown in table E-1. The imputation rates are calculated by dividing the number of missing responses by the number of persons who should have responded to the item; since skip patterns modify the interview universe for any given question, rates calculated on the entire sample universe would be misleading.

Some items are imputed because a respondent did not respond to the entire module (or wave interview); about 7 percent of those persons eligible for the School Enrollment and Financing module did not respond to any question in the module. (More than half of these were nonrespondents for the entire interview.) Despite the presence of the total module nonrespondents, most module questions are answered by most persons; of the 7,810 persons responding "yes" to the first item (the enrollment question), 66 percent had no imputed items in this section, and 87 percent had 2 or fewer imputations.

Table E-1. **Imputation and Edit Rates for Selected School Enrollment and Financing Items**

Item	Rate percent
Enrollment status[1]	7
Level of enrollment	4
Aid Recipiency[1]	31
Costs of schooling[2]	29-35
Lived away from home	14
Amount of aid received[3]	14-65

[1]These items have undergone extensive editing and allocation and have not been imputed.
[2]Includes rates for amount of tuition and fees, books and supplies, and room and board.
[3]Includes rates for amounts of each individual aid category.

It should be noted that the basic item of enrollment and the actual yes/no items for recipiency (e.g., did ... receive a Pell Grant) are not part of the hot deck imputation scheme. Instead, these items undergo an extensive edit process which checks information in other places in the questionnaire and previous interviews. As table E-1 shows, about 4 percent of the enrollment level responses were imputed. In general, the rates for the educational financing section are somewhat high. This is because many aid recipients are not imputed, but edited based on information given in other parts of the questionnaire or in a prior interview. Consequently, for many respondents, we know from other data that aid had been received during the past year and what kind it was. This leaves only the actual amount to be imputed resulting in the high levels of imputation shown in table E-2. The imputation rates for costs range from 29 to 35 percent.[1] It is also important to note that only about 43 percent of all answers of "yes" to the enrollment question were given by a self-respondent. Since this answer determines the subuniverse for the remaining questions, over half of the amounts data is being provided by someone other than the actual subject.

[1]These levels are similar to those obtained in previous waves where this module was administered.

Table E-2. **Comparison of Postsecondary Schooling Costs for Undergraduates between SIPP and Administrative Estimates[1]**

Cost	Administrative estimate	SIPP 1990 wave 5 estimates		
		Total	Self	Proxy
Tuition	$3,016	$1,876	$1,462	$2,295
Room and board	3,545	3,340	3,288	3,331
Books	-	344	303	390

- Represents zero.

[1]SIPP estimates are only for students enrolled in college years 1 through 4 for comparability to administrative data sources.

REASONABLENESS OF DATA

Another means of determining data quality is by comparison of the weighted survey estimates to other data, either from elsewhere in the questionnaire, a different survey, or administrative estimates. If editing, imputation and weighting procedures are properly applied, the final weighted data should compare favorably with other known estimates of the same phenomenon.

Enrollment

The initial question asks persons if they were enrolled in school anytime during the past year. The parenthetical expression instructs the interviewer to tell the respondent to include any regular school such as elementary, high school or college, or any vocational, technical or business school. Clearly, this is a very general question, and should elicit a large number of responses. In fact it does, yielding a weighted estimate of about 34.7 million persons. There is no administrative number which can provide a good basis for comparison. School enrollment is generally determined in a "snapshot" context, that is, as of a certain date what numbers of people were and were not enrolled in school. The October Current Population Survey (CPS), for instance, is the other basic Census tool for measuring school enrollment. Here, the item concerning enrollment is referenced to the interview week. Other surveys conducted by the Department of Education and the National Center for Education Statistics also use a "snapshot" approach in collecting, data. At levels beyond high school, enrollment may not be a year-long activity; people move in and out of the system much more rapidly. Consequently, estimates obtained from the snapshot approach should be lower than those yielded by a question such as the one used in SIPP. The point of closest correspondence should occur at the elementary and high school level, where fall enrollment numbers probably accurately reflect how many persons will be in those levels at any time during the year.

At the combined elementary and secondary level, the 1990 Wave 5 SIPP estimate of 13.0 million persons is about the same as the October CPS estimate of 13.1 million persons. The SIPP estimate is based on the number of persons who were age 15 or above during the summer of 1991 who were enrolled at the elementary and secondary levels at some point during the previous year. The CPS estimate is based on the number of students age 14 and above enrolled at the elementary and secondary levels (in October 1990) and removing from that total the approximate number of students, i.e. about one fourth of 14 year olds, who would not have turned 15 (the age of SIPP eligibility) before the time of the SIPP interview in summer 1991. This adjustment makes the population more comparable between the two surveys.

At the college level, the SIPP estimate of 16.8 million persons is higher than the October 1990 CPS estimate of 13.6 million. Using the Integrated Postsecondary Education Data System (IPEDS), Fall Enrollment Survey, the Department of Education estimated fall 1990 postsecondary enrollment to be 13.9 million. The SIPP estimate is larger than both the CPS and IPEDS estimate which would be expected since SIPP asks about school enrollment for any time within the last year, while the CPS reference period is only the previous week, and IPEDS is referenced in the fall only. Since college enrollment and non-regular schooling is not as likely as elementary and secondary to be year-round, the IPEDS estimate is expected to be lower even though it includes enrollment figures for all post-secondary schooling. The estimate for post-secondary schools other than college is estimated at 4.8 million in Wave 5 of the 1990 panel.

Educational Costs

The first amount items in the section ask questions regarding the costs of education, including tuition and fees, books and supplies and room and board for persons living away at school. Strictly comparable administrative figures are not available, but estimates for undergraduate college students from IPEDS probably provide the best administrative data. The IPEDS data come from the "Fall Enrollment" and the "Institutional Characteristics" surveys. Estimates of the mean tuition, room and board and books and supplies costs are shown in table E-2.

For the 1990-91 school year (the period most comparable to the SIPP period of reference for this module), the average tuition and fees were estimated to be $3,016. The 1990 SIPP Wave 5 estimate for persons in college years 1 through 4 is $1,876. The cost of room and board derived from the Department of Education data, was $3,545 a year; in SIPP, the estimate is $3,340. The estimate of the cost of books is $344, and there is no corresponding independent estimate for comparison.

Three contributing factors to the "underestimation" may be: 1) the high proportion of cases requiring imputation; 2) the fact that for many of the cases for

which "direct" data is received, it is taken from a proxy; and 3) greater representation of very short-term students (with lower costs) in the SIPP data. In fact, as table E2 shows, examination of tuition amounts by self/proxy status reveals that the average amounts reported by proxies (probably parents) is much closer to the derived administrative estimate than is the estimate taken as a self-report (that is, from the student themselves). In addition, the estimates are expected to be lower since Department of Education figures are estimated from institutions as year-round costs. SIPP averages are the means for each student for the past year; for many students the costs of the past year may include only one semester of tuition, thus lowering the average. These administrative estimates of tuition and fees are also weighted by full-time students only. SIPP estimates do not distinguish between full-time and part-time students.

Financial Aid Recipiency

The major data in this section are those concerning the receipt of educational financial aid and the amounts received from various sources. Respondents are able to report the receipt of 11 different types of financial aid as well as a twelfth residual "anything else" category. Some of the types of aid for which data is collected correspond closely to known financial aid programs, while others are of a more general nature. Table E-3 shows the comparison of some weighted SIPP estimates, both in terms of recipients and average amounts, to administrative data (where it is available).

With respect to the total number of recipients in specific programs, the general pattern of the data indicate that the SIPP estimates are close to some administrative and college board estimates. (As always, one should remember that these estimates may not be directly comparable in all cases to the reference period for the SIPP data.) However, some point estimates fall below other estimates, indicating that there is room for improvement. Part of the problem in collecting detailed sources such as these is that respondents may not be able to recall the specific program from which their funds came, especially when the report is given by a proxy. In this regard, the estimate for any specific program may not be very precise, but the overall estimate of all educational financing sources is probably much more comprehensively measured than in any single administrative context. Of course, that is what SIPP is supposed to be able to do—-measure the conjoint occurrence of different financial sources.

Examination of the dollar amounts reported by the recipients of these programs continues to show some discrepancies from the administrative and college board estimates (where available). While the mean amounts received for several programs correspond closely to the administrative numbers (note those for the Pell and GSL programs), some SIPP estimates are higher than the available administrative estimates. Unfortunately, for many sources of educational aid, comparative administrative data do not exist; thus it is not possible to determine if the estimates of sources such as "employer assistance" and "tuition reductions" are accurate.

The estimates of recipients and amounts for financial aid sources continue to show some variation from other available administrative estimates. The lack of exact knowledge and comparability of any and all external data sources we might find, however, should lead users to show caution in the detailed analysis of any specific kind of aid. Individuals using these data might instead draw their focus in terms of "total packages" of aid and costs; in this respect these data would seem to offer a high degree of reasonableness.

Table E-3. **Comparison of Aid Recipients and Amount of Aid Received Between SIPP and Administrative Estimates**

Source	Recipients[1]			Average amount received[2]		
	SIPP	College board [3]	Other administrative estimates[4]	SIPP	College board	Other administrative estimates
Pell Grant	3,047	3,300	3,405	$1,390	$1,489	$1,449
College Work Study	617	876	687	1,523	940	1,059
SEOG	420	678	761	1036	648	661
National Direct Student Loan	868	804	660	2,000	1,070	1,318
Guaranteed Student Loan	2,838	3,633	4,187[5]	2,870	2,709	2,804

[1]Numbers in thousands.
[2]Reported in current 1990 dollars.
[3]Data from the College Board are from "Trends in Students Aid: 1981 to 1991".
[4]Data are from the Department of Education: "Pell Grant: End of the Year Report," "Updated Tables and Graphs for the FY1991 Guaranteed Student Loan Data Book," and unpublished data sources.
[5]The number of Guaranteed Student Loan recipients is calculated as the number of guaranteed loans divided by 1.15 (the average number of loans per student, as reported by Department of Education).

DATA FROM THE NATIONAL POSTSECONDARY STUDENT AID STUDY (NPSAS)

Users who are familiar with the Department of Education's NPSAS data may notice discrepancies between the NPSAS and SIPP estimates. Although these two surveys are both nationally representative samples, the universes differ and as a result estimates may also differ. Although these two surveys reflect two different academic years, 1989-90 for NPSAS and 1990-91 for SIPP, there should be some correspondence. Table E-4 provides an indication of how the populations differ between the two surveys.[2]

Table E-4. **Number of Students Enrolled by Level of Enrollment**

(Numbers in thousands)

	Level of enrollment			
	Total	Undergraduate (2 and 4-year institutions)	Other undergraduate	Graduate
SIPP90	20,560	12,380	4,203	3,977
Dependent	6,149	5,412	560	176
Percent	30	44	13	4
Independent	14,410	6,967	3,642	3,801
Percent	70	56	87	96
NPSAS89-90	18,590	14,879	1,391	2,318
Dependent[1]	7,846	7,367	391	87
Percent	42	50	28	4
Independent	10,679	7,464	983	2,231
Percent	57	50	71	96

[1]Since 65,500 weighted cases were unclassified, NPSAS numbers do not add to total.

In NPSAS, students are characterized by academic level, undergraduate and graduate (identified through institutional records), and by institutional type. For this table, undergraduates were divided into two groups, undergraduates in 2-year and 4-year colleges and those in "less than 2-year" institutions. In SIPP, students are self-identified by actual enrollment level (college years 1 through 6+ and vocational, technical, business, or other type of postsecondary school). These students were classified as follows: 1) college years 1 through 4 as undergraduates in 2-year and 4-year colleges; 2) vocational, business, technical, and other institutions as undergraduates in a less than 2-year institution; and 3) college years 5 and higher as graduate students. Although these categories are not exactly comparable, they do provide interesting findings. The SIPP data clearly show a greater enrollment in the "other undergraduate" institutions than does NPSAS. This is most likely due to the ability of SIPP to collect data for those students of the shortest enrollment durations—usually in nontraditional postsecondary institutions. Why would there be more short-term students captured in SIPP? Institutions are ineligible in NPSAS if they offer only correspondence courses; offer only courses or seminars of less than three months duration; or provide only avocational, recreational, or remedial courses. Students in courses of less than 3 months duration and the other types of courses mentioned are very likely to have reported themselves as enrolled in the SIPP survey since the enrollment question is so broad. On a different level, the number of SIPP graduate students may be higher than in NPSAS since students are classified by enrollment level. Fifth-year undergraduates may be included in this rough categorization of graduate students in SIPP, while in NPSAS, students are identified by actual type of program.

Upon further examination, it is clear that the differences in the enrollment numbers may lead to different estimates in average costs for groups of students. For example, the SIPP estimate of tuition and fees for those in other undergraduate institutions is $759, far below the NPSAS average of $4,123.[3] Again, this underestimate points to the differences in counting students of the shortest enrollment periods. Enrollment in a course for 1 month is likely to be much less in cost than a student enrolled for 6 months. The inclusion of nearly 3 million more students may certainly drive down the cost average, if, as suspected, these students are those of very short enrollment durations. Furthermore, table E-4 indicates that these missing students are more likely to be independent students who tend to have lower costs than dependent students (see table 2 of report). These non-traditional students may also be more likely to be considered "less than half-time" students. Although SIPP, does not differentiate between full-time and part-time students, unpublished NPSAS data indicates that tuition and fees drop dramatically depending on attendance status (full-time students average $3,332; at least half-time, but less than full-time students average $1,110; and less than half-time students average $596 in tuition and fees).

A comparison of undergraduates in 2-year and 4-year colleges is more difficult to make. The NPSAS data clearly indicate that students enrolled in 2-year colleges have substantially lower tuition and fees (only $854) than do those undergraduate students in 4-year colleges ($3,199 for non-PhD-granting schools and $3,380 for PhD-granting schools). The SIPP estimate cannot reliably estimate the cost for students in 2-year versus

[2]The weighted NPSAS estimates can be found in a technical report from the National center for Education Statistics entitled "Methodology Report for the 1990 National Postsecondary Student Aid Study." The estimates are found in the executive summary of the report.

[3]The NPSAS data on average costs are from unpublished data provided form the National Center for Education Statistics.

4-year institutions as data for type of institution is not available. The SIPP estimates show that undergraduates enrolled in the first 2 years of college are have lower tuition and fees than those in the 3rd and 4th years ($1,667 vs. $2,179) indicating that the inclusion of 2-year college undergraduates has driven down the number. However, it is impossible to disaggregate the groups to make a true comparison of this level of students.

SUMMARY

While the educational financing data collected in the 5th Wave of the 1990 panel of SIPP appears to have a high degree of reasonableness and utility, there are important differences from the other sources of financial aid data of which users should be aware. For example, estimates of the number of recipients and the amounts they receive for specific aid sources show some variability from the available administrative estimates. Caution should, therefore, be exercised in detailed analysis of specific aid sources; however, in terms of "overall" pictures of students, their costs and their sources of aid, the data as a whole appear reasonable. Variation from other data, such as the NPSAS survey, may be a function of the inclusion of a large component of very short-term students in the SIPP data. Without additional variables for disaggregation in the SIPP, however, analytic comparability of universes between the two surveys is not possible.

Several additional points should be kept in mind when using these data: 1) Edits/Imputations The implementation (in the 1985 Panel) of a more rigorous edit procedure which checks data from both the core and three prior waves to look for the actual report of any of the aid sources identified in the topical module seems to have worked quite well. Nevertheless, this increase in the number of "inferred" recipiencies provides a large base for the number of cases which must then have an amount imputed. This explains imputation rates of around 50 percent for some specific amount sources; 2) Proxy Responses - Probably because of the nature of the subpopulation of concern (i.e., students away at school), proxy response is quite high for the enrollment and financial aid items. This in turn acts to drive up the nonresponse (and imputation) rate, particularly for items which do not have closed-ended response categories, and items which require an amount as a response. Additionally, for items such as tuition and room and board costs, proxy responses seem to be much closer to administrative estimates than those given as self-reports. One possibility is that the proxies (parents) have a better idea of the amounts they may be paying than do the students, many of whom are not responsible for paying the bills. Much of the financial aid, however, may go directly to the institution and thus is never really seen by the respondent, whether self- or proxy-interview; 3) Amounts - In general, the ability of an individual to return a reliable amount (or any amount), even for self-respondents, is less than the ability to return a yes-no or closed-ended response. The simple item non-response rates of amount items versus other types of items demonstrates this point.